Active learning for deep semantic parsing

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, Mark Johnson
Voicebox Technologies
{longd,hadia,dominiquee,glenp,philipc,markj}@voicebox.com

Abstract

Semantic parsing requires training data that is expensive and slow to collect. We apply active learning to both traditional and "overnight" data collection approaches. We show that it is possible to obtain good training hyperparameters from seed data which is only a small fraction of the full dataset. We show that uncertainty sampling based on least confidence score is competitive in traditional data collection but not applicable for overnight collection. We evaluate several active learning strategies for overnight data collection and show that different example selection strategies per domain perform best.

1 Introduction

Semantic parsing maps a natural language query to a logical form (LF) (Zettlemoyer and Collins, 2005, 2007; Haas and Riezler, 2016; Kwiatkowksi et al., 2010). Producing training data for semantic parsing is slow and costly. Active learning is effective in reducing costly data requirements for many NLP tasks. In this work, we apply active learning to deep semantic parsing and show that we can substantially reduce the data required to achieve state-of-the-art results.

There are two main methods for generating semantic parsing training data. The traditional approach first generates the input natural language utterances and then labels them with output LFs. We show that active learning based on uncertainty sampling works well for this approach.

The "overnight" annotation approach (Wang et al., 2015) generates output LFs from a grammar, and uses crowd workers to paraphrase these LFs into input natural language queries. This approach is faster and cheaper than traditional annotation. However, the difficulty and cost of data generation and validation are still substantial if we need a large amount of data for the system to achieve high accuracy; if the logical forms can express complex combinations of semantic primitives that must be covered; or if the target language is one with relatively few crowd workers.

Applying active learning to the overnight approach is even more compelling, since the unlabelled LFs can be generated essentially for free by a grammar. However, conventional active learning strategies are not compatible with the overnight approach, since the crowd annotators produce inputs (utterances) rather than labels (LFs).

In order to apply active learning to deep semantic parsing, we need a way of selecting hyperparameters without requiring the full training dataset. For optimal performance, we should rerun hyperparameter tuning for each active learning round, but this is prohibitively expensive computationally. We show that hyperparameters selected using a random subset of the data (about 20%) perform almost as well as those from the full set.

Our contributions are (1) a simple hyperparameter selection technique for active learning applied to semantic parsing, and (2) straightforward active learning strategies for both traditional and overnight data collection that significantly reduce data annotation requirements. To the best of our knowledge we are the first to investigate active learning for overnight data collection.

2 Related work

Sequence-to-sequence models are currently the state-of-the-art for semantic parsing (Jia and Liang, 2016; Dong and Lapata, 2016; Duong et al., 2017). In this paper, we also exploit a sequence-to-sequence model to minimise the amount of la-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 43–48
Melbourne, Australia, July 15 - 20, 2018.

belled training data required to achieve state-of-the-art semantic parsing results.

Active learning has been applied to a variety of machine learning and NLP tasks (Thompson et al., 1999; Tang et al., 2002; Chenguang Wang, 2017) employing various algorithms such as least confidence score (Culotta and McCallum, 2005), large margin (Settles and Craven, 2008), entropy based sampling, density weighting method (Settles, 2012), and reinforcement learning (Fang et al., 2017). Nevertheless, there has been limited work applying active learning for deep semantic parsing with the exception of Iyer et al. (2017). Different from conventional active learning, they used crowd workers to select what data to annotate for traditional semantic parsing data collection.

In this paper, we apply active learning for both traditional and overnight data collection with the focus on overnight approach. In addition, a limitation of prior active learning work is that the hyperparameters are usually predefined in some way, mostly from different work on the same or similar dataset, or from the authors experience (Wang et al., 2017; Fang et al., 2017). In this paper, we investigate how to efficiently set the hyperparameters for the active learning process.

3 Base S2S Model

We base our approach on the attentional sequence-to-sequence model (S2S) of Bahdanau et al. (2014). This attentional model uses a bidirectional recurrent neural network (RNN) to encode a source as a sequence of vectors, which are used by another RNN to generates output. Given the source utterance $x = [x_1, x_2, ...x_n]$ and target LF $y = [y_1, y_2, ...y_m]$, we train the model to minimize the loss under model parameters θ.

$$\texttt{loss} = -\sum_{i=1}^{m} \log \mathsf{P}(y_i|y_1, ..y_{i-1}, x; \theta) \quad (1)$$

Additionally, we apply the UNK replacement technique in Duong et al. (2017), keeping the original sentence in the data.[1]

4 Active learning models

There is a diversity of strategies for active learning. A simple and effective active learning strategy is based on **least confidence score** (Culotta and McCallum, 2005). This strategy selects utterance x' to label from the unlabelled data U_x as follows:

$$x' = \underset{x \in U_x}{\operatorname{argmin}} \left[\max_{y^*} \mathsf{P}(y^*|x; \theta) \right]$$

where y^* is the most likely output. We found that this least confidence score works well across datasets, even better than more complicated strategies in traditional data collection (described below).

4.1 Traditional data collection

In the traditional (forward) approach, we start with the list of unlabelled utterances and an initial seed of utterances paired with LFs. We gradually select utterances to annotate with the aim of maximizing the test score as early as possible. We use forward S2S sentence loss as defined in Equation (1) as the least confidence score measurement (i.e. select the instance with higher loss).

The drawback of a least confidence score strategy (and strategies based on other measurements such as large margin), is that they only leverage a single measurement to select utterances (Settles and Craven, 2008). To combine multiple measurements, we build a classifier to predict if the model will wrongly generate the LF given the utterance, and select those utterances for annotation. The classifier is trained on the data generated by running 5-fold cross validation on annotated data.[2] We exploit various features, including sentence log loss, the margin between the best and second best solutions, source sentence frequency, source encoder last hidden state and target decoder last hidden state (see supplementary material §A.1 for more detail) and various classifier architectures including logistic regression, feedforward networks and multilayer convolutional neural networks. On the development corpus, we observed that the least confidence score works as well as the classifier strategy.

4.2 Overnight data collection

In the overnight (backward) approach, we start with the set of all unlabelled LFs (U_y), and an initial randomly-selected seed of LFs paired with utterances (i.e. labelled LFs L_y). The aim is to select

[1] We call S2S model applied to traditional data collection and overnight data collection as forward S2S and backward S2S respectively. The forward S2S model estimates $\mathsf{P}(y|x)$, the backward S2S model estimates $\mathsf{P}(x|y)$.

[2] This classifier is complementary to the approach proposed in Iyer et al. (2017) where we use this classifier instead of user feedback.

LFs for which we should obtain utterances, maximizing the test score as early as possible. In the overnight approach, we can't use the least confidence score (i.e. the forward S2S sentence loss) directly since we can't estimate $\mathsf{P}(y|x)$ because we don't know the utterance x. We have to somehow approximate this probability with regard to the performance on test.

A simple strategy is just to apply the backward S2S model and estimate $\mathsf{P}(x|y)$, e.g. we select LF y' to label from the unlabelled data U_y as follows:

$$y' = \underset{y \in U_y}{\operatorname{argmin}} \left[\max_{x^*} \mathsf{P}(x^*|y;\theta) \right]$$

Essentially, we train the S2S model to predict the utterance given the LF. The motivation is that if we can reconstruct the utterance from the LF then we could possibly generate LFs from utterances. However, this strategy ignores one important aspect of semantic parsing, which is that LFs are an abstraction of utterances. One utterance is mapped to only one LF, but one LF corresponds to many utterances.

Since the forward S2S loss performs so well, another strategy is to approximate the selections made by this score. We train a linear binary classifier[3] to predict selections, using features which can be computed from LFs only. We extract two set of features from the LF model and the backward S2S model. The LF model is an RNN language model but trained on LFs (Zaremba et al., 2014).[4] We extract the LF sentence log probability i.e. $\log \mathsf{P}(y)$, feature from this model. The backward S2S model, as mentioned above, is the model trained to predict an utterance given a LF. We extracted the same set of features as mentioned in §4.1 including LF sentence log loss, margin between best and second best solutions, and LF frequencies.

On the development corpus, we first run one active learning round using forward S2S model sentence loss (i.e. modelling $\mathsf{P}(y|x)$) on the initial annotated data L_y. The set of selected LFs based on forward S2S loss will be the positive examples, and all other LFs that are not selected will be the negative examples for training the binary classifier. Our experiments show that the classifier which uses the combination of two features (source LF frequencies and the margin of best and second best solution) are the best predictor of what is selected by forward S2S model log loss (i.e. modelling $\mathsf{P}(y|x)$). It is interesting to see that absolute score of backward S2S model loss is not a good indicator as it is not selected. This may be due to the fact that utterance-LF mapping is one-to-many and the model probability is distributed to all valid output utterances. Hence, low probability is not necessary an indicator of bad prediction. We use the linear combination of the two features mentioned above with the weights from the binary classifier as a means of selecting the LF for overnight active learning on different corpora without retraining the classifier.

5 Experiment

5.1 Datasets

We experiment with the NLMaps corpus (Haas and Riezler, 2016) which was collected using the traditional approach. We tokenize following Kočiský et al. (2016). We also experiment with the Social Network corpus from the Overnight dataset (Wang et al., 2015) (which was collected using the overnight approach). Social Network was chosen as being the largest dataset available. Since neither corpora have a separate development set, we use 10% of the training set as development data for early stopping. We select ATIS (Zettlemoyer and Collins, 2007) as our development corpus for all feature selection and experiments with classifiers in §4.1 and §4.2.

For evaluation, we use full LF exact match accuracy for all experiments (Kočiský et al., 2016). Note that this is a much stricter evaluation compared with running through database evaluator as in Wang et al. (2015).

5.2 Hyperparameter tuning

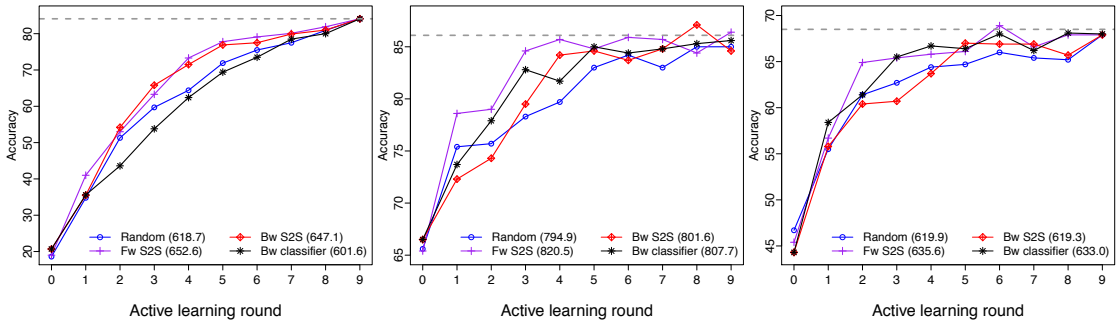
Hyperparameter tuning is important for good performance. We tune the base S2S model (§3) on the development data by generating 100 configurations using Adam optimizer (Kingma and Ba, 2014) and a permutation of different source and target RNN sizes, RNN cell types, initializer, dropout rates and mini-batch sizes.

As mentioned, hyperparameter tuning is often overlooked in active learning. The common approach is just to use the configuration from a similar problem, from prior work on the same dataset,

[3] Instead of binary classifier, it would also be possible to train a logistic model. However, we leave this for future work.

[4] We use the configuration from Zaremba et al. (2014).

[5] The exact match accuracy for Social Network is extracted from logs from (Jia and Liang, 2016).



(a) NLMap (b) ATIS (c) Overnight (Social Network)

Figure 1: Active learning for various selection criteria. *Random* baseline randomly select the training data at each round. *Fw S2S* is used for traditional data collection using forward S2S loss score. *Bw S2S* is used for overnight data collection using backward S2S loss score. *Bw classifier* is used also for the overnight approach but linearly combines several scores together as mentioned in §4.2. The scores in parentheses measure the area under the curve. The dashed lines are the SOTA from Table 1.

	NLMap	Social	ATIS
From ATIS	76.0	65.8	86.0
Small subset	84.2	68.9	85.7
Full data	84.2	69.1	86.0
SOTA	84.1	68.8	86.1

Table 1: The LF exact match accuracy on NLMap, Social Network and ATIS with configurations from ATIS, from hyperparameter tuning on small subset of data (10% + dev) or on the full training data. The supervised SOTA for NLMap and ATIS (Duong et al., 2017) and Social Network (Jia and Liang, 2016) are provided for reference.[5]

or based on the authors own experience. However, in practice we don't have any prior work to copy the configuration from. Table 1 shows the experiments with the NLMap and Social Network corpora with configurations: 1) copied from another dataset (ATIS), 2) tuned on a small subset (10% of train data plus development data) and 3) tuned on the full dataset. We can see that copying from a different dataset results in a suboptimal solution, which is expected since the different datasets are significantly different. It is surprising that tuning on small subset of the data performs as well as tuning on all the data and, more importantly, it achieves similar results as the state of the art (SOTA).

5.3 Active Learning Results

Figure 1 shows the active learning curve for NLMap, ATIS and Overnight (Social Network) datasets. 10% of data is randomly selected as initial seed data for active learning and hyperparameter tuning. We run active learning for 10 rounds, selecting 10% of the data at each round. Round 0 reports the result trained on the initial seed data and round 9 is the result on the whole training data. For reference, we also report *Fw S2S* for Social Network, treating that corpus as if they were collected using the traditional approach, and *Bw S2S/classifier* for NLMap and ATIS treating those corpora as if they were collected using the overnight approach.

For traditional data collection (forward direction), S2S loss consistently outperforms the random baselines on both datasets. The differences are as high as 9% for NLMap (at round 4). Applying this strategy for ATIS, we reach SOTA results at round 4, using only 50% of data. We also experimented with the large margin baseline and classifier strategies as mentioned in §4.1. The least confidence strategy using S2S loss outperforms large margin and achieves similar performance with the more complicated classifier strategy, thus we omit those results for brevity.

On the overnight data collection active learning (backward direction), the results are split. The backward S2S loss performs particularly well on the NLMap corpus, approximating the forward S2S performance. However, it performs similar to the random baseline in the other corpora. On the other hand, the classifier strategy performs

well on both ATIS and Social Network but poorly on NLMap. Using this strategy, we approximate the SOTA for both ATIS and Social Network at round 5 and 6 respectively (saving 40% and 30% of data). We suspect that backward S2S loss performs so well on NLMap since there is a one-to-one mapping between utterance and LF. The number of unique LFs in the training data for NLMap, ATIS and Overnight are 95.4%, 28.4% and 19.5% respectively. All in all, our proposed strategies for "overnight" active learning are nearly as good as traditional active learning, showing in similar area under the curve value in Figure 1.

6 Conclusion

We have discussed practical active learning for deep semantic parsing. We have empirically shown that it is possible to get good hyperparameters from only a small subset of annotated data. We applied active learning for both traditional and overnight semantic parsing data collection. For traditional data collection, we show that least confidence score based on S2S log loss performs well across datasets. Applying active learning for overnight data collection is challenging, and the best performing strategy depends on the domain. We recommend that applications explore both the backward S2S and classifier strategies.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Laura Chiticariu Yunyao Li Chenguang Wang. 2017. Active learning for black-box semantic role labeling with neural factors. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. pages 2908–2914. https://doi.org/10.24963/ijcai.2017/405.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*. pages 746–751.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 33–43.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, pages 379–389. https://doi.org/10.18653/v1/K17-1038.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 595–605. http://aclweb.org/anthology/D17-1063.

Carolin Haas and Stefan Riezler. 2016. A corpus and semantic parser for multilingual natural language querying of openstreetmap. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 740–750.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. *CoRR* abs/1704.08760. http://arxiv.org/abs/1704.08760.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 12–22.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Tomáš Kočiský, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. 2016. Semantic parsing with semi-supervised sequential autoencoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1078–1087.

Tom Kwiatkowksi, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1223–1233.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 1070–1079.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 120–127. https://doi.org/10.3115/1073083.1073105.

Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '99, pages 406–414. http://dl.acm.org/citation.cfm?id=645528.657614.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. 2017. Cost-effective active learning for deep image classification. *CoRR* abs/1701.03551. http://arxiv.org/abs/1701.03551.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1332–1342.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR* abs/1409.2329. http://arxiv.org/abs/1409.2329.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. pages 678–687.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured /classification with probabilistic categorial grammars. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005*. pages 658–666.